In Silico Data
Miracles by the KY-methods

SS002-3

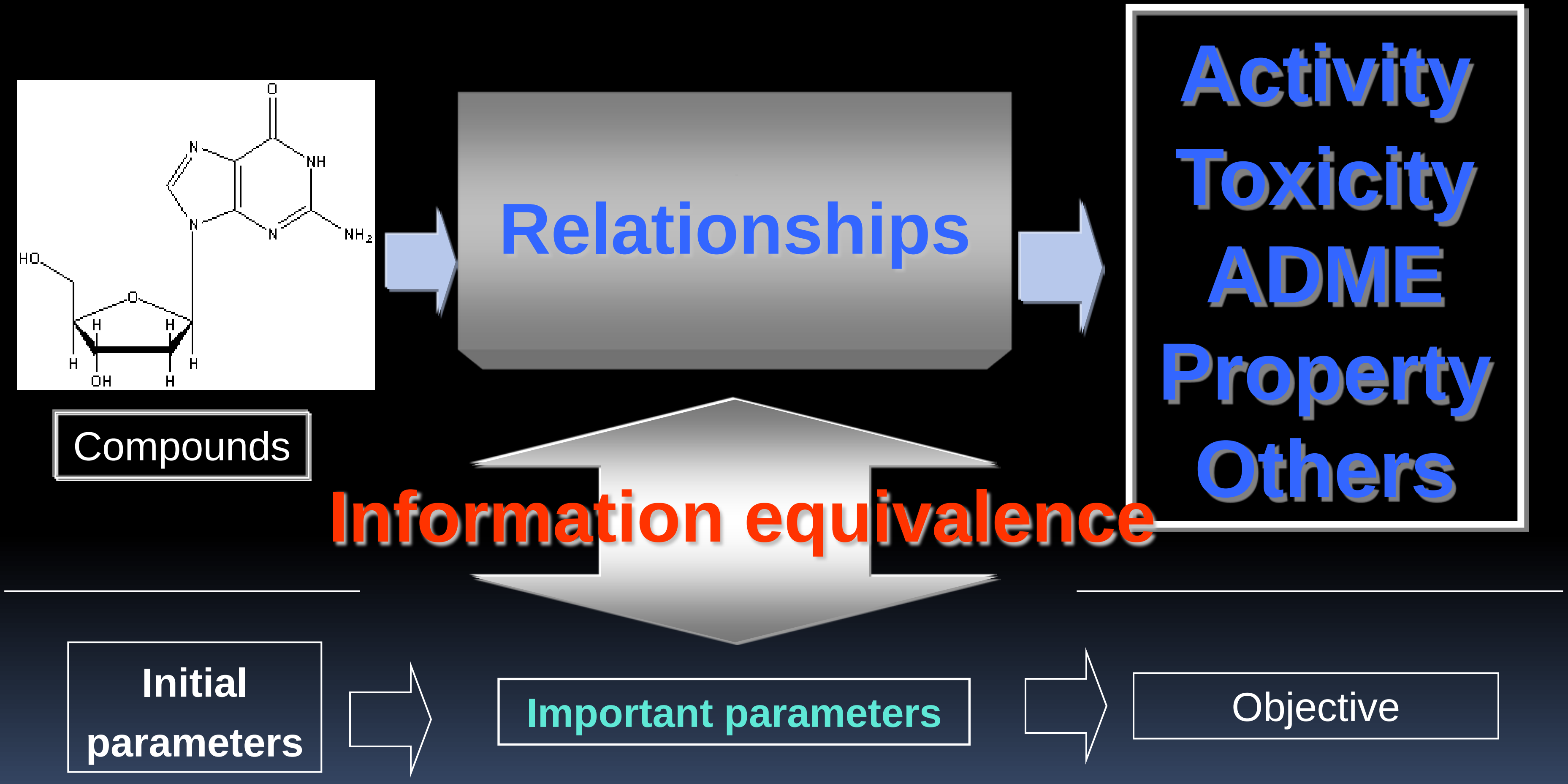
Qualitative structure-toxicity relationships (QSTR) on skin sensitization

ICOH Cancun Mexico 2012

Kohtaro Yuta1, Kazuhiro Sato2, Yukinori Kusaka2
1. In Silico Data Ltd., Chiba, Japan,
2.Department of Environmental Health, School of Medicine, University of Fukui, Fukui, Japan

1. Basic concept of QSTR approach
2. Sample and parameter handling
3. Data analysis and results by discriminant analysis
4. The KY-methods and conclusions

Basic concept of data analysis by multi-variate analysis and pattern recognition techniques



Compounds

Relationships

Activity
Toxicity
ADME
Property
Others

Information equivalence

Initial parameters

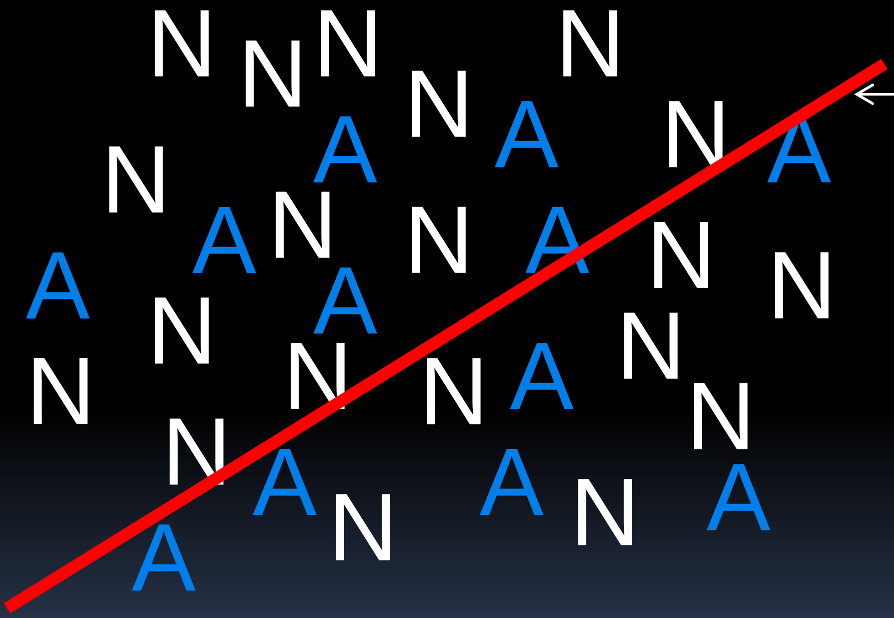
Important parameters

Objective

Execute data analysis methods

Relations between pattern space and analytical objects

N-dimensional pattern space including noisy data



Linear discriminant

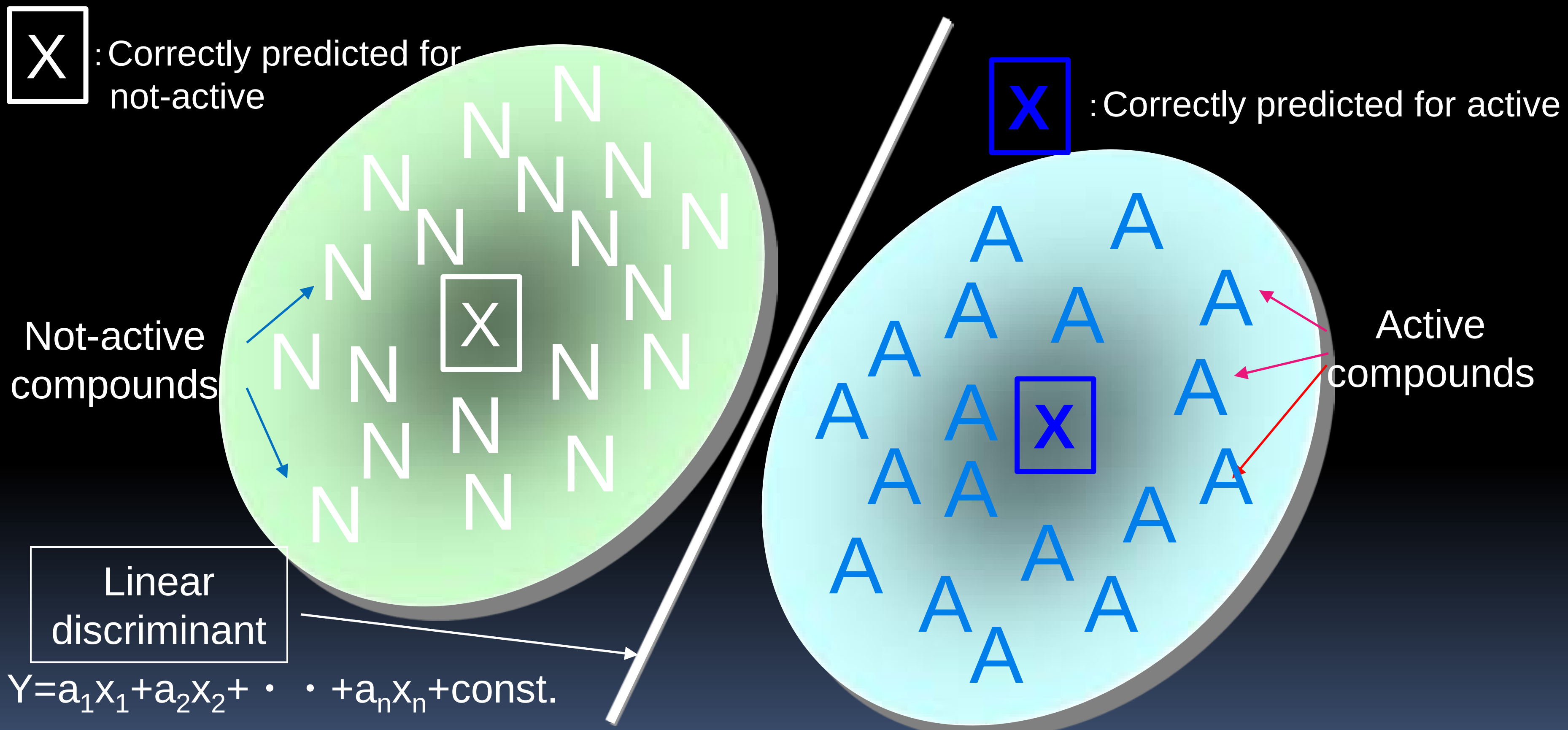
$Y=a_1x_1+a_2x_2+\cdot\ \ \cdot+a_nx_n+\text{const.}$

A; Active compounds
N; Not active compounds

No relations between pattern space and objects

Relations between pattern space and analytical objects

N-dimensional pattern space by intrinsic parameters → Prediction



X : Correctly predicted for not-active

X : Correctly predicted for active

Not-active compounds

Active compounds

Linear discriminant

$Y=a_1x_1+a_2x_2+\cdot\ \cdot+a_nx_n+\text{const.}$

1. Pattern space divided into active and not-active compounds
2. This pattern space is classified by linear discriminant function

Style of discriminant function and regression equation

$$Y = +/- \; a_1x_1 \; +/- \; a_2x_2 \; +/- \; \cdot \; \cdot \; \cdot \; +/- \; a_nx_n \; +/- \; \text{const.}$$

Y : activity, ADME, toxicity, property

$Y >= 0$

active or toxic

$Y < 0$

not active or non-toxic

Analysis of activity, ADME, toxicity or property

Coefficient ai >= 0
parameter Xi

･go up activity and toxicity

Coefficient ai < 0
parameter Xi

･go down activity and toxicity

Structure-activity and Structure-toxicity relationships

1. Basic flow of QSAR approach

2. **Sample and parameter handling**

3. Data analysis and results by discriminant analysis

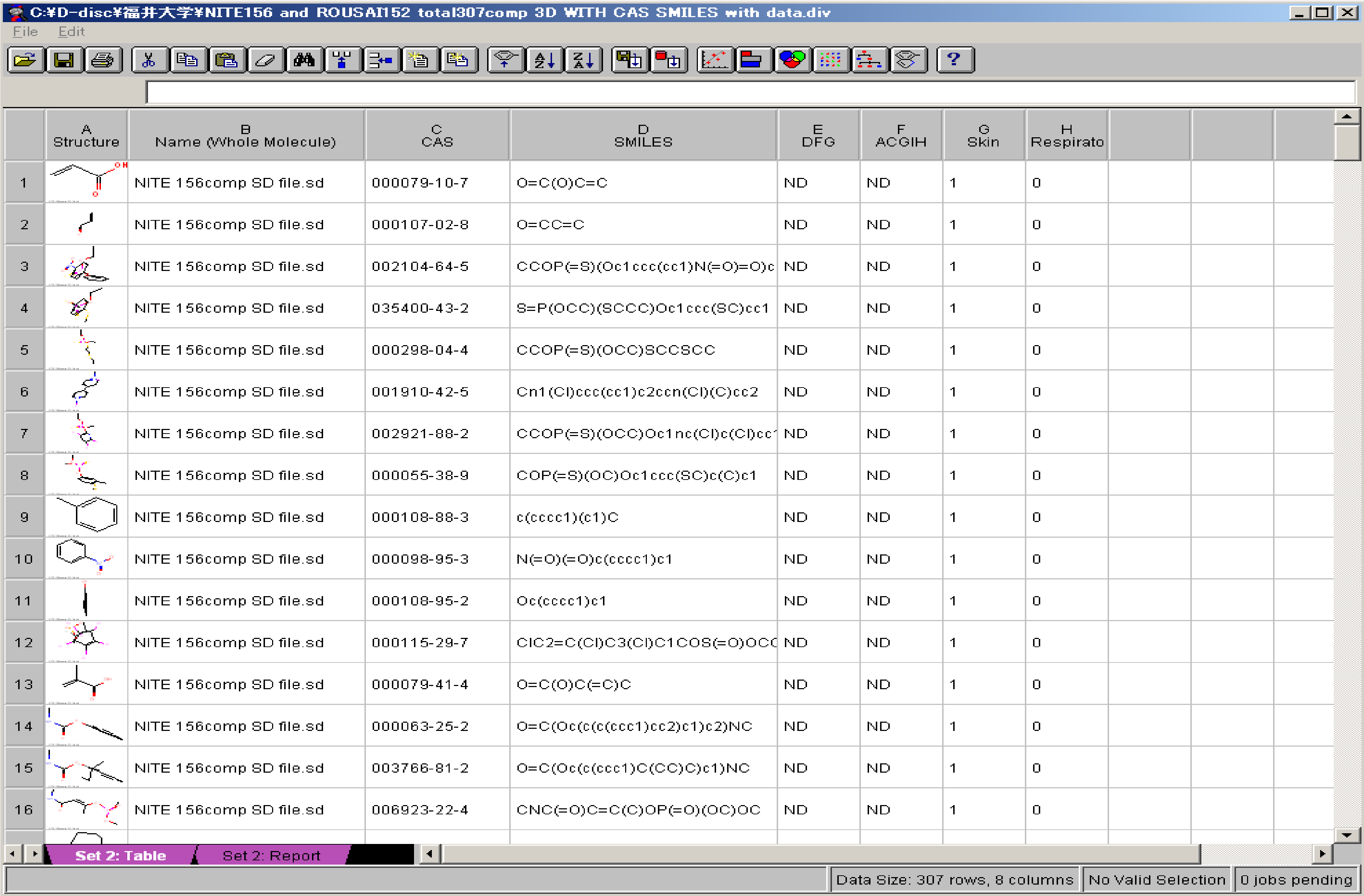
4. The KY-methods and conclusions

Used samples :
obtained from following 4 different databases

1. Maximale Arbeitsplatz-Konzentration (MAK)
   ····＞ positive skin sensitizer
2. Biologischer Arbeitsstoff-Toleranz-Wert (BAT)
   ····＞ positive skin sensitizer
3. Deutschen Forschungsgemeinschaft (DFG)
   ····＞ positive skin sensitizer
4. Japanese Globally Harmonized System of Classification and Labeling of Chemicals (GHS) Inter-ministerial Committee of the National Institute for Technology and Evaluation ····＞ negative skin sensitizer

Total **593** compounds
419 positive skin sensitizer
174 negative skin sensitizer

List of the used samples
(Structure, CAS number, SMILES code, Sensitization)

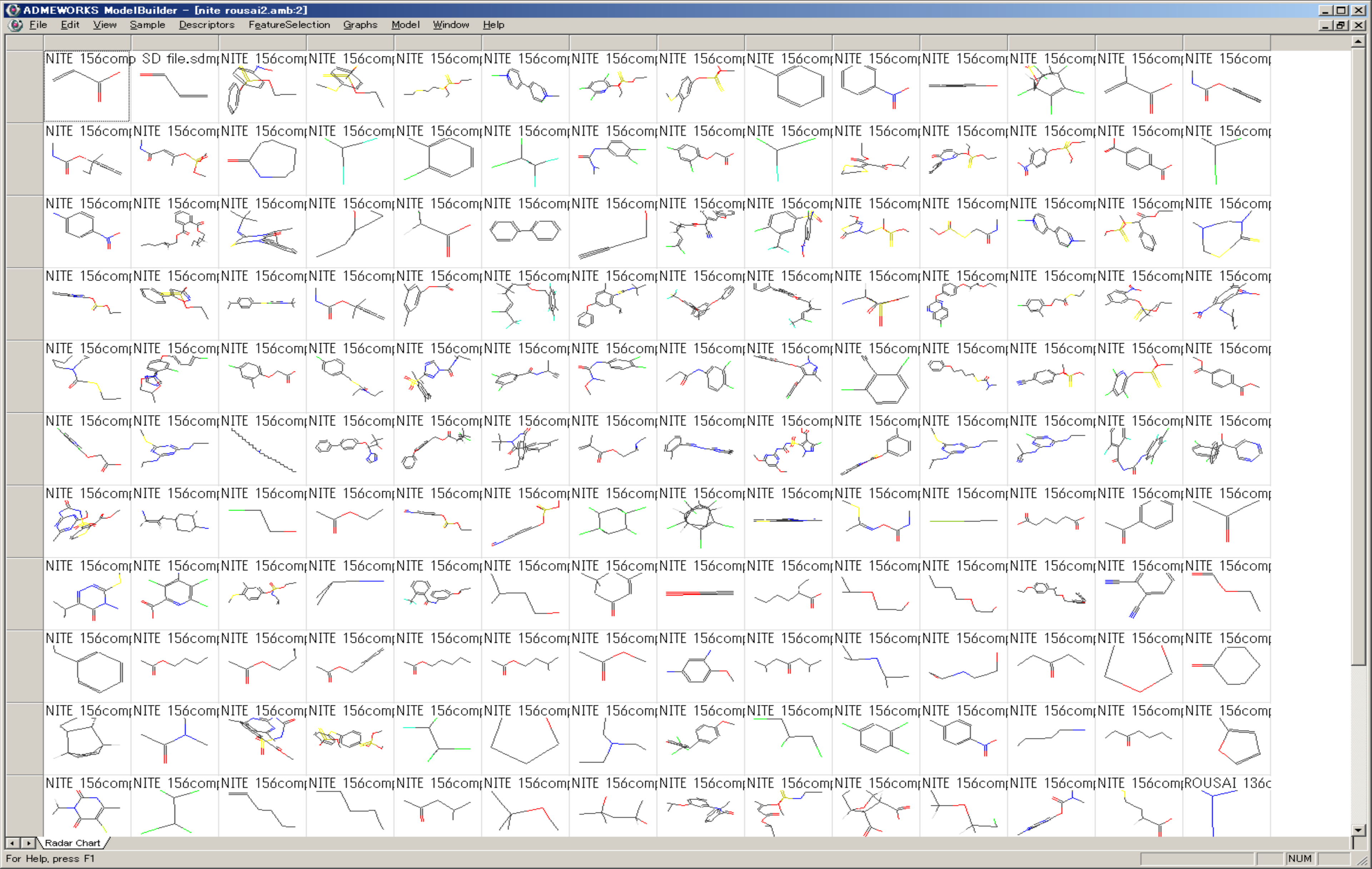


C:¥D-disc¥福井大学¥NITE156 and ROUSAI152 total307comp 3D WITH CAS SMILES with data.div

	A Structure	B Name (Whole Molecule)	C CAS	D SMILES	E DFG	F ACGIH	G Skin	H Respirato
1		NITE 156comp SD file.sd	000079-10-7	O=C(O)C=C	ND	ND	1	0
2		NITE 156comp SD file.sd	000107-02-8	O=CC=C	ND	ND	1	0
3		NITE 156comp SD file.sd	002104-64-5	CCOP(=S)(Oc1ccc(cc1)N(=O)=O)c	ND	ND	1	0
4		NITE 156comp SD file.sd	035400-43-2	S=P(OCC)(SCCC)Oc1ccc(SC)cc1	ND	ND	1	0
5		NITE 156comp SD file.sd	000298-04-4	CCOP(=S)(OCC)SCCSCC	ND	ND	1	0
6		NITE 156comp SD file.sd	001910-42-5	Cn1(Cl)ccc(cc1)c2ccn(Cl)(C)cc2	ND	ND	1	0
7		NITE 156comp SD file.sd	002921-88-2	CCOP(=S)(OCC)Oc1nc(Cl)c(Cl)cc	ND	ND	1	0
8		NITE 156comp SD file.sd	000055-38-9	COP(=S)(OC)Oc1ccc(SC)c(C)c1	ND	ND	1	0
9		NITE 156comp SD file.sd	000108-88-3	c(cccc1)(c1)C	ND	ND	1	0
10		NITE 156comp SD file.sd	000098-95-3	N(=O)(=O)c(cccc1)c1	ND	ND	1	0
11		NITE 156comp SD file.sd	000108-95-2	Oc(cccc1)c1	ND	ND	1	0
12		NITE 156comp SD file.sd	000115-29-7	ClC2=C(Cl)C3(Cl)C1COS(=O)OCC	ND	ND	1	0
13		NITE 156comp SD file.sd	000079-41-4	O=C(O)C(=C)C	ND	ND	1	0
14		NITE 156comp SD file.sd	000063-25-2	O=C(Oc(c(c(ccc1)cc2)c1)c2)NC	ND	ND	1	0
15		NITE 156comp SD file.sd	003766-81-2	O=C(Oc(c(ccc1)C(CC)C)c1)NC	ND	ND	1	0
16		NITE 156comp SD file.sd	006923-22-4	CNC(=O)C=C(C)OP(=O)(OC)OC	ND	ND	1	0

Data Size: 307 rows, 8 columns | No Valid Selection | 0 jobs pending

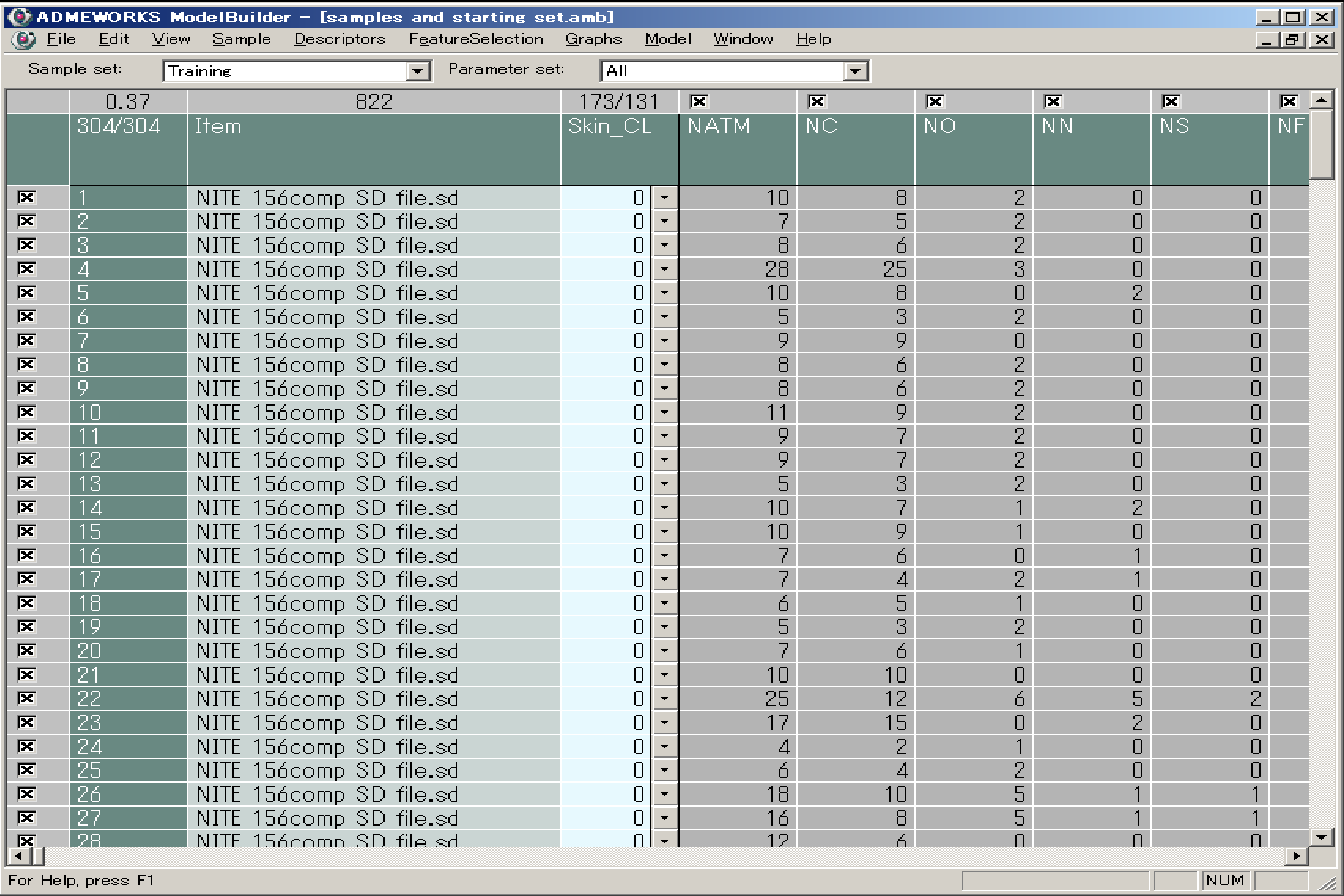
List of 3-Dimensional Structures of the used compounds



NITE 156comp SD file.sdm		NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp
NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	NITE 156comp	ROUSAI 136c

For Help, press F1

List of Compounds and Generated 822 Parameters



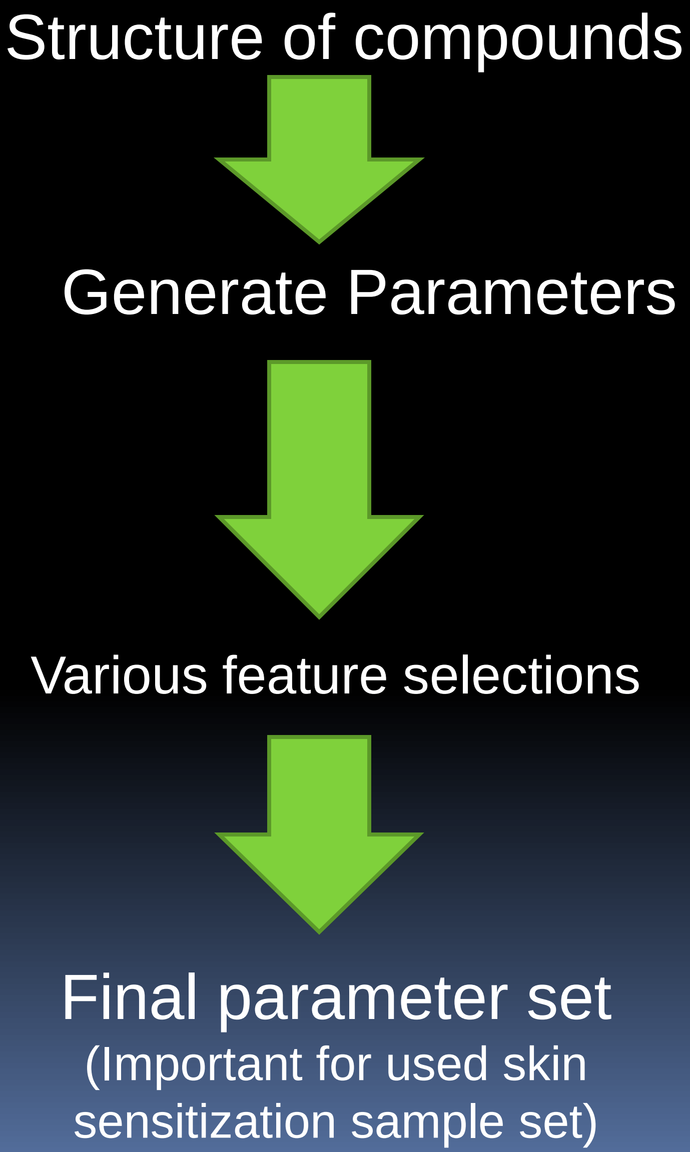
Sample set: Training

Parameter set: All

	0.37	822	173/131	☒	☒	☒	☒	☒	☒
	304/304	Item	Skin_CL	NATM	NC	NO	NN	NS	NF
☒	1	NITE 156comp SD file.sd	0	10	8	2	0	0	
☒	2	NITE 156comp SD file.sd	0	7	5	2	0	0	
☒	3	NITE 156comp SD file.sd	0	8	6	2	0	0	
☒	4	NITE 156comp SD file.sd	0	28	25	3	0	0	
☒	5	NITE 156comp SD file.sd	0	10	8	0	2	0	
☒	6	NITE 156comp SD file.sd	0	5	3	2	0	0	
☒	7	NITE 156comp SD file.sd	0	9	9	0	0	0	
☒	8	NITE 156comp SD file.sd	0	8	6	2	0	0	
☒	9	NITE 156comp SD file.sd	0	8	6	2	0	0	
☒	10	NITE 156comp SD file.sd	0	11	9	2	0	0	
☒	11	NITE 156comp SD file.sd	0	9	7	2	0	0	
☒	12	NITE 156comp SD file.sd	0	9	7	2	0	0	
☒	13	NITE 156comp SD file.sd	0	5	3	2	0	0	
☒	14	NITE 156comp SD file.sd	0	10	7	1	2	0	
☒	15	NITE 156comp SD file.sd	0	10	9	1	0	0	
☒	16	NITE 156comp SD file.sd	0	7	6	0	1	0	
☒	17	NITE 156comp SD file.sd	0	7	4	2	1	0	
☒	18	NITE 156comp SD file.sd	0	6	5	1	0	0	
☒	19	NITE 156comp SD file.sd	0	5	3	2	0	0	
☒	20	NITE 156comp SD file.sd	0	7	6	1	0	0	
☒	21	NITE 156comp SD file.sd	0	10	10	0	0	0	
☒	22	NITE 156comp SD file.sd	0	25	12	6	5	2	
☒	23	NITE 156comp SD file.sd	0	17	15	0	2	0	
☒	24	NITE 156comp SD file.sd	0	4	2	1	0	0	
☒	25	NITE 156comp SD file.sd	0	6	4	2	0	0	
☒	26	NITE 156comp SD file.sd	0	18	10	5	1	1	
☒	27	NITE 156comp SD file.sd	0	16	8	5	1	1	
☒	28	NITE 156comp SD file.sd	0	12	6	0	0	0	

For Help, press F1

822 parameter generation from structure of compound and final 60 parameter set after feature selection process



Structure of compounds

Generate Parameters

Various feature selections

Final parameter set

(Important for used skin sensitization sample set)

Total **593** compounds

419 positive skin sensitizer

174 negative skin sensitizer

Total **822** parameters per compound

- topological (2-D) parameters
  MC parameters, etc..
- topographical (3-D) parameters
  Box parameters, etc..
- property parameters
  LogP, MR, Volume, Surface, etc..
- electric parameters
  HOMO, LUMO, etc..
- substructure parameters
  Count of substructures, etc..

Final **60** parameters

1. Basic flow of QSAR approach

2. Sample and parameter handling

3. **Data analysis and results by discriminant analysis**

4. The KY-methods and conclusions

List of Classification results by Various Discriminant Analysis

1. **NN (Neural Network)** :60 Parameter set

   Classification ratio : 85.5%

2. **Linear discriminant analysis by least squares algorithms**

   Classification ratio : 85.7%

3. **SVM(Support Vector Machine)** :60 Parameter set

   Classification ratio : 90.7%

4. **ADA Boost:** 60 Parameter set

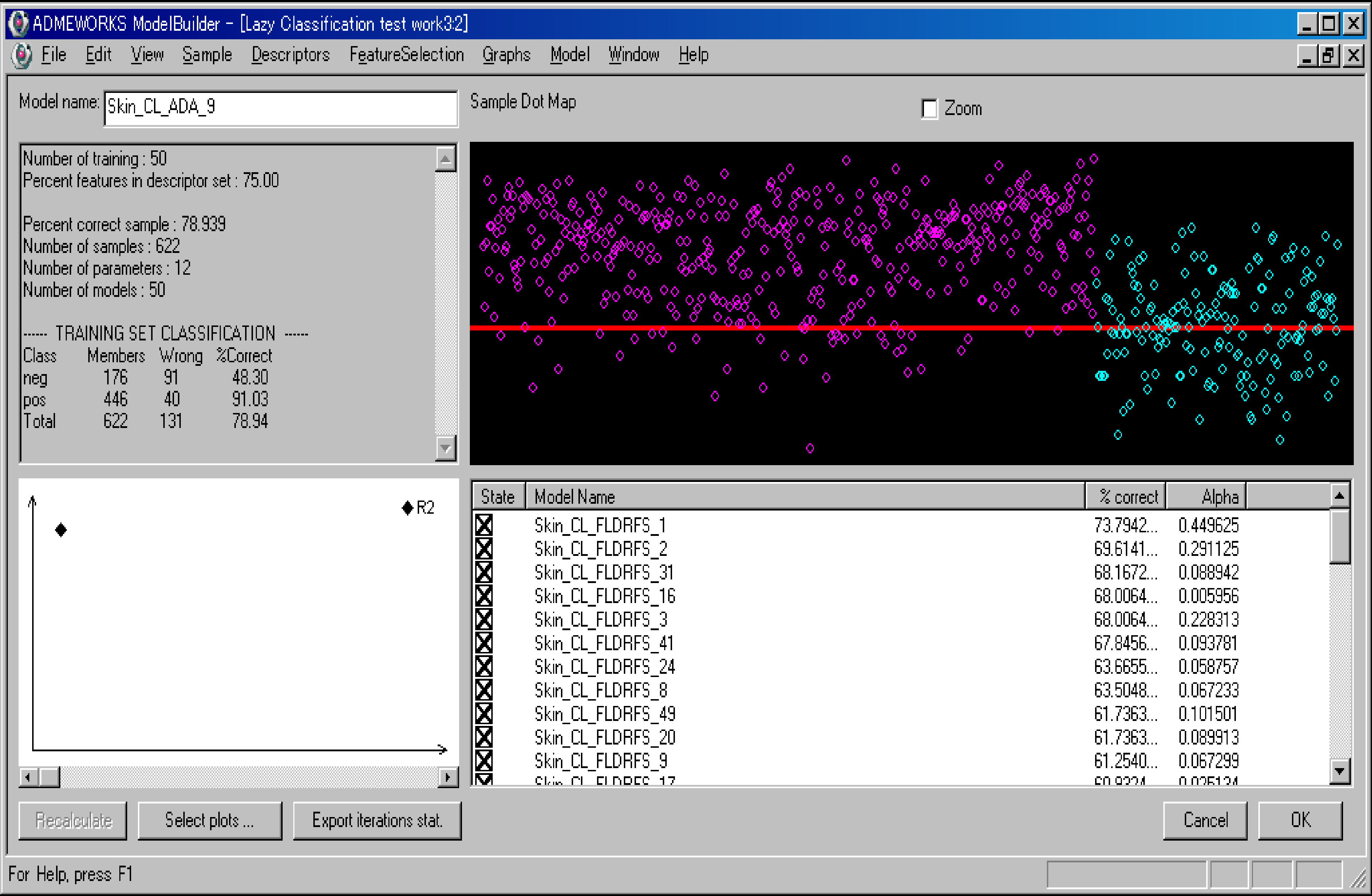
   Classification ratio : 77.5%

5. KY-method for Discriminant analysis

Classification racio : 100% (Perfect Classification)

Perfect classification

Incomplete classification example by the AdaBoost (77.5%)



Model name: Skin_CL_ADA_9

```
Number of training : 50
Percent features in descriptor set : 75.00

Percent correct sample : 78.939
Number of samples : 622
Number of parameters : 12
Number of models : 50

----- TRAINING SET CLASSIFICATION -----
Class      Members   Wrong   %Correct
neg          176       91      48.30
pos          446       40      91.03
Total        622      131      78.94
```

State	Model Name	% correct	Alpha
☒	Skin_CL_FLDRFS_1	73.7942...	0.449625
☒	Skin_CL_FLDRFS_2	69.6141...	0.291125
☒	Skin_CL_FLDRFS_31	68.1672...	0.088942
☒	Skin_CL_FLDRFS_16	68.0064...	0.005956
☒	Skin_CL_FLDRFS_3	68.0064...	0.228313
☒	Skin_CL_FLDRFS_41	67.8456...	0.093781
☒	Skin_CL_FLDRFS_24	63.6655...	0.058757
☒	Skin_CL_FLDRFS_8	63.5048...	0.067233
☒	Skin_CL_FLDRFS_49	61.7363...	0.101501
☒	Skin_CL_FLDRFS_20	61.7363...	0.089913
☒	Skin_CL_FLDRFS_9	61.2540...	0.067299
☒	Skin_CL_FLDRFS_17	60.9324...	0.025124

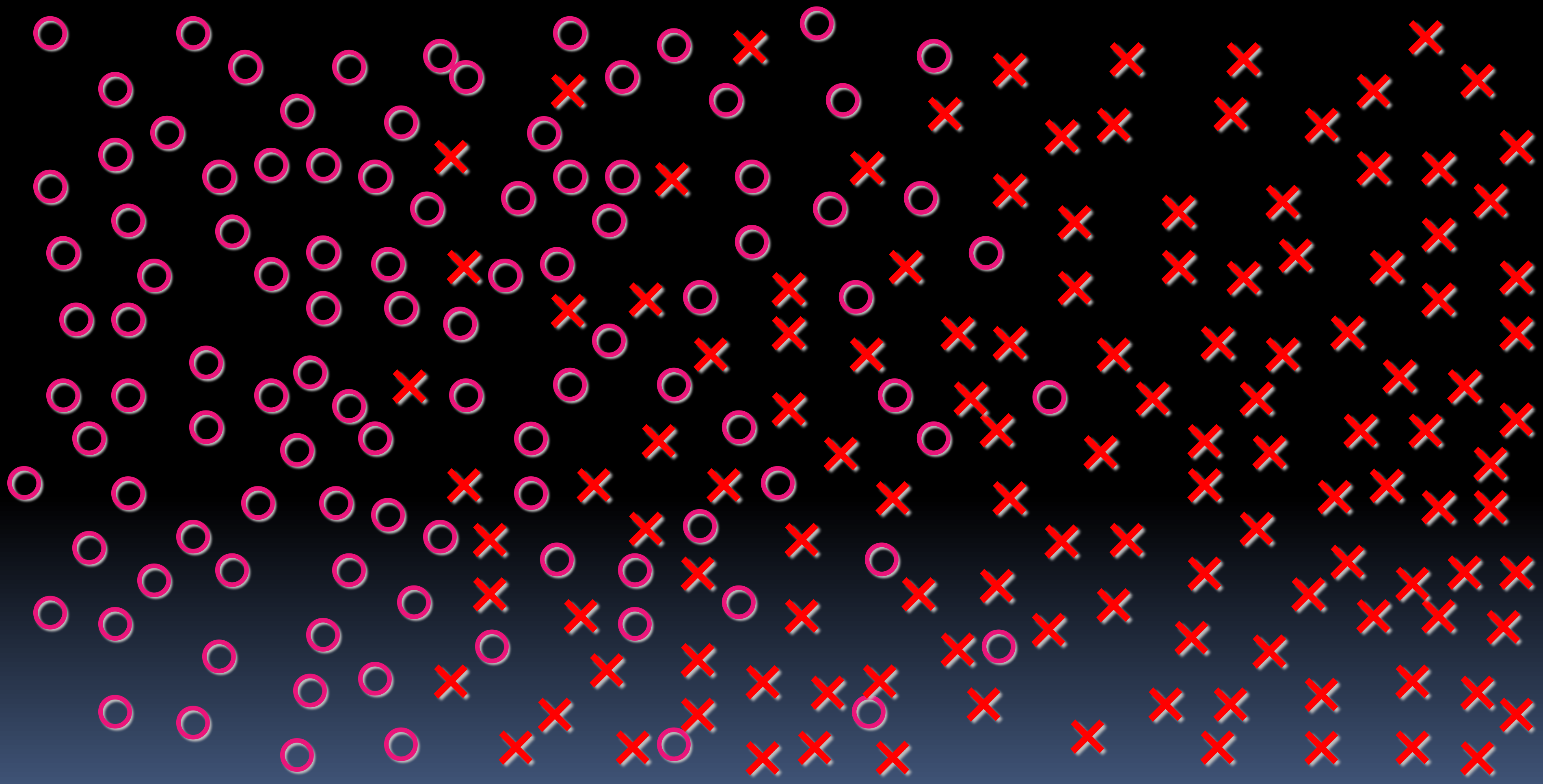
1. Basic flow of QSAR approach

2. Sample and parameter handling

3. Data analysis and results by discriminant analysis

4. The KY-methods and conclusions

Sample space : Highly overlapped space

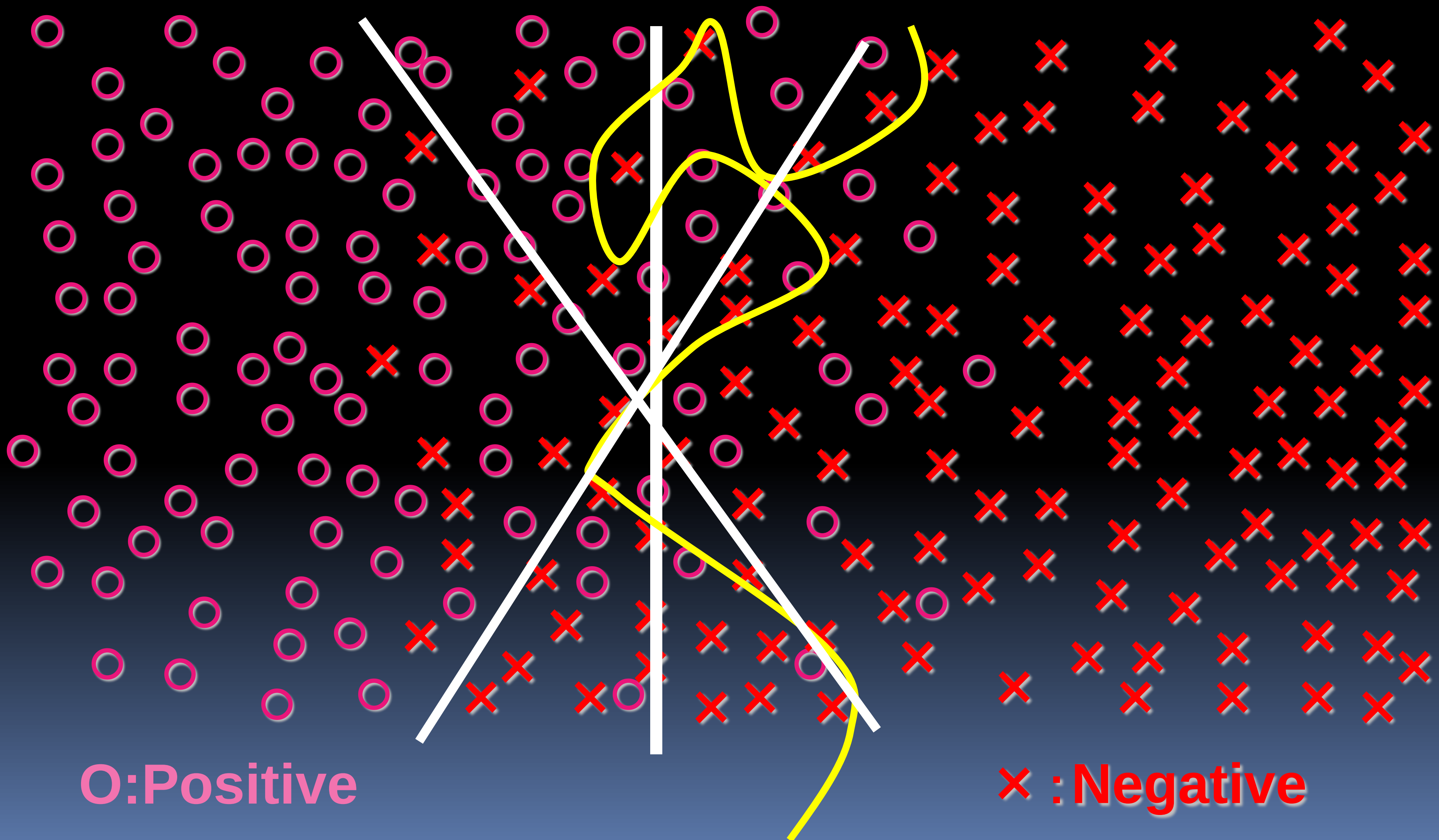


O:Positive

× :Negative

Sample space : Highly overlapped space

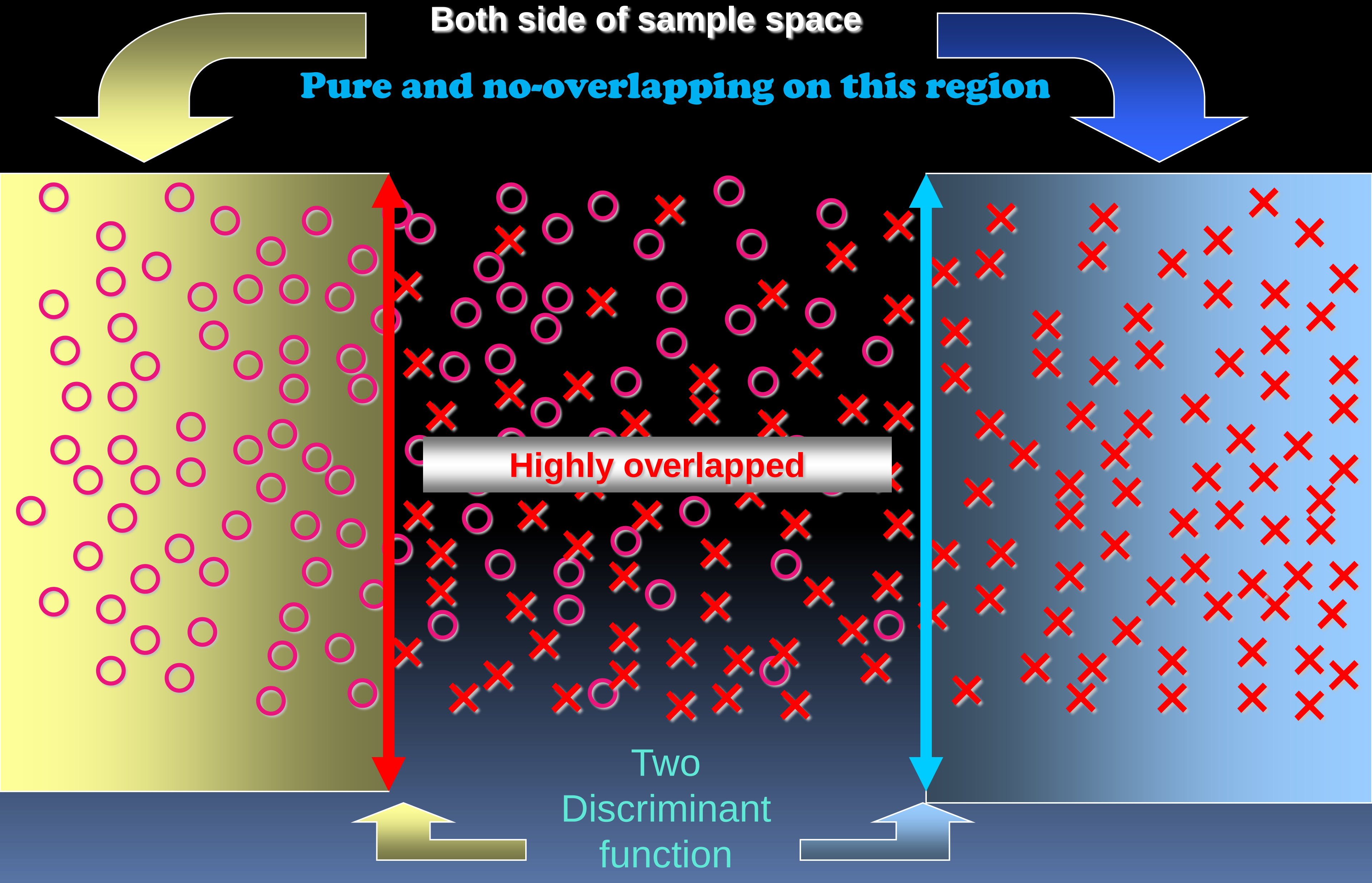
Discriminant function : Linear and non-linear



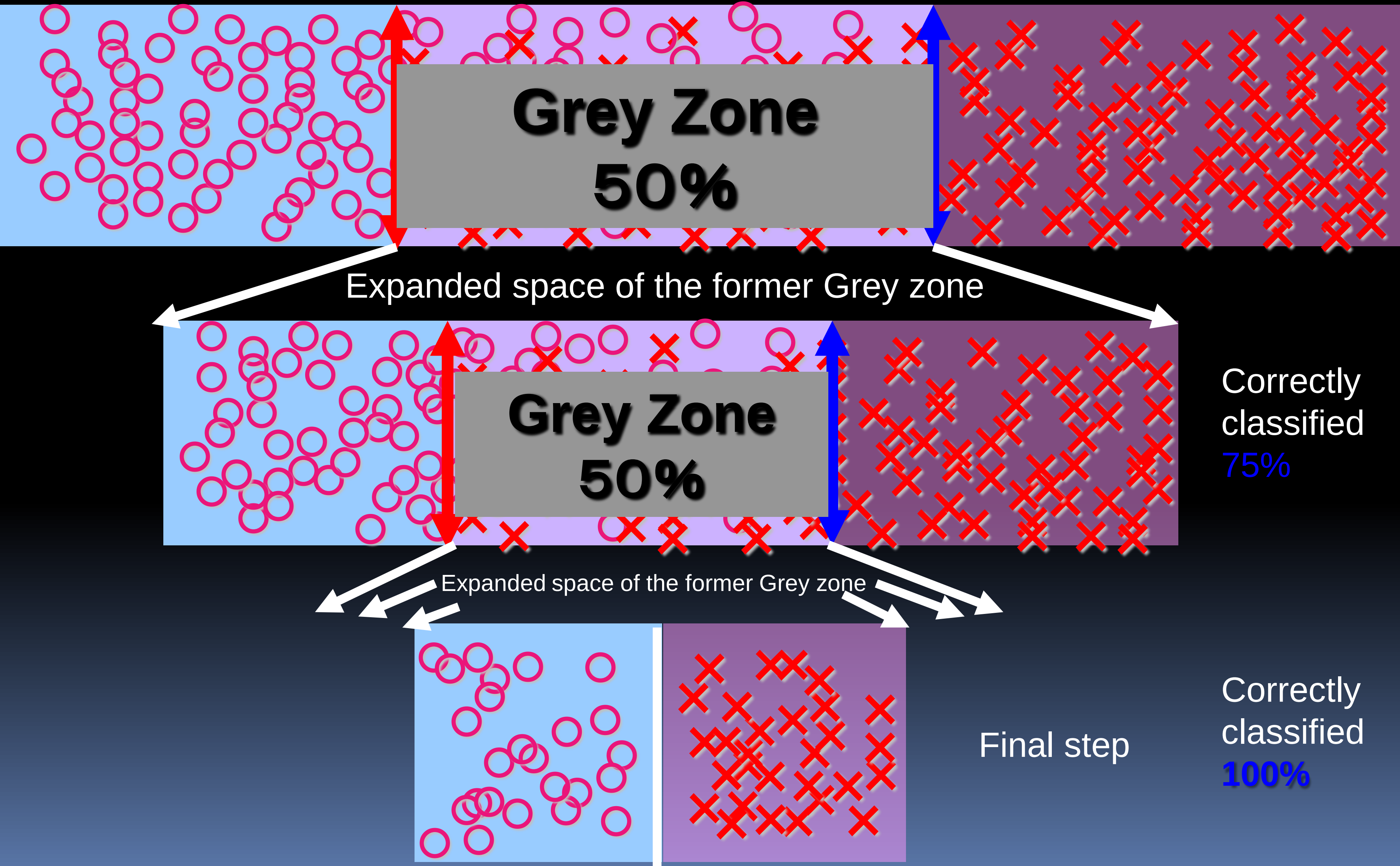
O:Positive

× :Negative

Spatial region on sample space


Both side of sample space
Pure and no-overlapping on this region
Highly overlapped
Two Discriminant function

Steps to the K-step methods



Grey Zone
50%

Expanded space of the former Grey zone

Grey Zone
50%

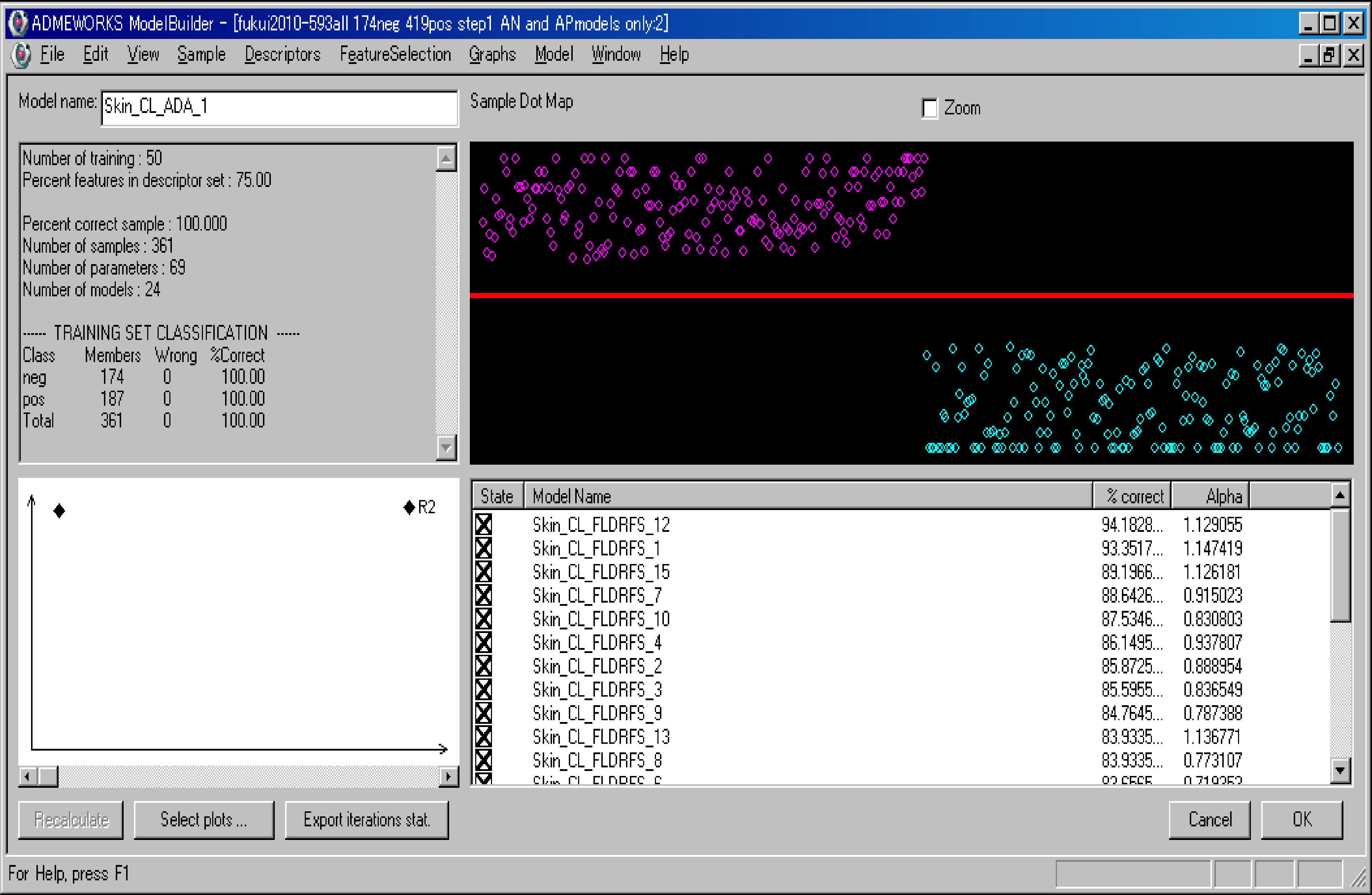
Correctly classified
75%

Expanded space of the former Grey zone

Final step

Correctly classified
100%

Perfect classification example by the KY-method (Displayed by the AdaBoost)



Model name: Skin_CL_ADA_1

```
Number of training : 50
Percent features in descriptor set : 75.00

Percent correct sample : 100.000
Number of samples : 361
Number of parameters : 69
Number of models : 24

----- TRAINING SET CLASSIFICATION -----
Class     Members   Wrong   %Correct
neg         174       0      100.00
pos         187       0      100.00
Total       361       0      100.00
```

State	Model Name	% correct	Alpha
☒	Skin_CL_FLDRFS_12	94.1828...	1.129055
☒	Skin_CL_FLDRFS_1	93.3517...	1.147419
☒	Skin_CL_FLDRFS_15	89.1966...	1.126181
☒	Skin_CL_FLDRFS_7	88.6426...	0.915023
☒	Skin_CL_FLDRFS_10	87.5346...	0.830803
☒	Skin_CL_FLDRFS_4	86.1495...	0.937807
☒	Skin_CL_FLDRFS_2	85.8725...	0.888954
☒	Skin_CL_FLDRFS_3	85.5955...	0.836549
☒	Skin_CL_FLDRFS_9	84.7645...	0.787388
☒	Skin_CL_FLDRFS_13	83.9335...	1.136771
☒	Skin_CL_FLDRFS_8	83.9335...	0.773107
☒	Skin_CL_FLDRFS_6	83.5555...	0.719252

Spatial features of the “KY-methods”

Always achieve perfect classification

(a) Even if **the number of samples becomes very large**,
the KY-methods achieves perfect (100%) classification

(b) Even if **overlapped sample space grows too big**,
the KY-methods achieves perfect (100%) classification

Differences between the KY-methods and the ordinal methods

1. Number of classified sample zone:
   KY-methods ; three zones   Ordinal methods ; two zones

2. Repeat number of classification:
   KY-methods ; > = 2 times   Ordinal methods ; 1 time

Patented: US 7,725,413   Patent pended: Japan, Korea, EU

SS002-3

Thank you for your kind attention

ICOH Cancun Mexico 2012

Kohtaro Yuta
In Silico Data Ltd.
E-Mail : k-yuta@insilicodata.com
http://www.insilicodata.com